Analysis of oligo hybridization properties by high-resolution tiling microarrays in rice

Xiangfeng Wang, Lei Li, Viktor Stolc, Waraporn Tongprasit, Chen Chen, Jun Wang, Songgang Li, and Xing Wang Deng

Rice genome sequencing and computational annotation provide a static map for understanding this model of *Gramineae* species. With the development of *in situ* oligonucleotide synthesis technology, tiling-path microarrays have become a dynamic and efficient way for monitoring large-scale transcriptional activities and detecting novel transcribed elements missed by software. Unlike conventional cDNA or oligonucleotide arrays, tiling-path platforms employ the full extent of oligos covering given genomic regions, and thus offer excellent experimental conditions in which to assay the properties of oligos in terms of their specificity and efficiency of hybridization to their corresponding targets. Here, we report a tiling-path microarray analysis of a 1-Mb region (10 to 11 Mb) in japonica rice chromosome 10, which was tiled by a 36-mer oligo set at a resolution of 5 bp. Our analysis focused on three major factors of oligo hybridization properties, including GC content, melting temperature (Tm), and the repetitiveness of oligo sequences.

Keywords: Rice, genomics, tiling-path microarrays, transcriptome, hybridization

Whole-genome tiling-path microarrays have been used in several sequenced model organisms as a dynamic and efficient way to facilitate *in silico* genome annotations. Traditional cDNA or oligomer microarrays use relatively few probes for each gene and are biased toward known and predicted gene structures (Mockler et al 2005). In contrast, tiling-path platforms provide more biologically relevant information beyond the measurement of mRNA levels. This includes genome-wide transcriptional activity monitoring, gene structure definition, the identification of novel transcribed elements, and broader uses in organ-specific alternatively spliced isoforms, chromatin-immunoprecipitation-chip studies, and other extended applications, according to different innovatively customized designs (Bertone et al 2005).

Characteristics of the rice genome and its annotation

Rice is one of the most important crops in the world, providing staple food for about half of the human population (Hoshikawa 1993). Determination of the rice genome sequence has been considered a milestone in the field of *Gramineae* species, and a valuable resource for scientists to investigate its functional elements. The genome-scale comparison of the two subspecies of rice, indica and japonica, respectively, sequenced by the Beijing Institute of Genomics and the International Rice Genome Sequencing Project, provides us with important clues in deducing the evolutionary history of grass species. Studies of single nucleotide polymorphisms between indica and japonica and their related wild species are equally important for understanding heterosis, or hybrid vigor, in rice.

Genomics analyses in rice discovered that many of rice's own unique features were different from those of other sequenced organisms. For instance, unlike most other grasses, rice has a relatively small genome of about 440 Mb (Bennetzen et al 2002), but the initial estimate of 55,000 to 60,000 rice genes is nearly twice as many as the gene content of mammalian genomes, and rice genes are usually clustered together into islands, separated by highly repetitive sequences. However, recent published estimates based on newly improved genome data indicate that the nontransposon gene count is at least 38,000 to 40,000 (Yu et al 2005). In the IRGSP newly released finished quality sequence of japonica, 37,544 protein-coding genes were identified, of which 71% had a putative homolog in *Arabidopsis*. In a reciprocal analysis, 90% of the *Arabidopsis* proteins had a putative homolog in the predicted rice proteome (International Rice Genome Sequencing Project 2005). It is possible that most of the low-homology genes might be the remnants of ancient gene duplications or fragments resulting from the transposition of transposable elements. The effort to identify pseudogenes in rice needs more evidence of expression and comparative genomics methodology.

One of the unique characteristics of the rice gene is the gradient change in GC content along the transcription direction (Wong et al 2002). This feature and the unusual high-GC content represent the major unique factors that might affect the accuracy of microarray experiments in rice. The other crucial factor that might affect these experiments is the repetitive sequences presenting the transcripts. In rice, the repetitive sequences account for nearly 45% of the genome, composed of thousands of retrotransposons, and numerous miniature inverted-repeat transposable element copies. In a recent publication on the improved rice indica genome sequence, Yu et al (2005) reported an ancient whole-genome duplication covering 65.7% of the current genome that can be dated back to a common time before the divergence of grasses. They also identified 18 distinct segmental duplication pairs and massive ongoing individual gene duplications that provide a never-ending source of raw material for gene genesis, and are major contributors to the differences between members of the grass family (Yu et al 2005).

Tiling-path microarray analysis in rice

We have been taking bioinformatics approaches and experimental methods centered on tiling-path microarray to facilitate and improve the rice genome annotation. The tiling-path strategy was first used on rice japonica chromosome 4, in which a total of 15,242 nonredundant polymerase chain reaction (PCR)-amplified genomic fragments were printed on slides as probes and hybridized with RNA samples pooled from six representative rice organs. Our analysis revealed a chromatin-level regulation of both protein-coding genes and transposable element–related genes, in hetero- and euchromatin regions at different developmental stages (Jiao et al 2005).

We also used the Maskless Array Synthesizer (MAS) platform to design higher-resolution tiling experiments, by using 36-mer oligos to cover all the nonrepetitive sequences of the newly assembled indica and japonica genomes. The implementation of this design generated about 6.5 and 6.1 million probe pairs for tiling approximately 60% of the regions (nonrepetitive portions) of the indica and japonica genome, leading to the resolution of each 36-mer oligo interrogating every 60-bp genomic sequence on average. The two sets of 34 and 32 arrays were hybridized with mixed cDNA targets derived from four tissues: seedling root, seedling shoot, panicle, and suspension-cultured cells.

To examine rice tiling-array hybridization conditions, and to establish a data-processing procedure, the pilot analysis was first conducted on two sets of hybridization data representing indica and japonica chromosome 10, both consisting of nearly 3,000 nontransposon gene models. More than 80% of the annotated gene models were verified and nearly 500 novel intergenic transcribed regions were detected by our tiling-array platform. The overview of global signals along chromosome 10 indicates that expression of transcriptome can be related to chromosomal architecture (Li et al 2005).

Using tiling microarrays to experimentally analyze probe hybridization properties

Tiling-path microarray provides an unbiased observation of genome-wide transcription activity and is a powerful approach complementary to computer-based annotation methods. Unlike the traditional cDNA or oligomer microarray, in which probes are preselected to adapt to the calculated optimal parameters and specificities, tiling-path platforms employ the full extent of oligos covering given genomic regions, and thus offer excellent experimental conditions in which to assay the properties of oligos in terms of their specificity and efficiency of hybridization to their corresponding targets. In particular, the following questions can be addressed:

1. The diverse properties of probes produce unequal hybridization intensities despite being located in the same transcribed region. For instance, oligos with high GC content are more easily hybridized with targets than those with low GC content, but they also might introduce false signals if GC composition is beyond a certain percentage. Likewise, oligos within a same gene might vary a lot in hybridization efficiency because of their different

melting temperatures. Therefore, the diversity of probe property is one of the reasonable interpretations for a large proportion of gene models lacking or having unequal hybridization intensities.

2. In rice, duplication events are common phenomena and many remain in the current genome. Although duplicated pairs diverged in both function and sequence during evolution, they still share a similar DNA composition. Cross-hybridization is therefore unavoidable in most tiling-array experiments and biases the accuracy of determining whether a gene is expressed or not.
3. As a common procedure in oligo design, RepeatMasker is generally used to mask highly repeated sequences to avoid redundant oligos being selected, but this procedure inevitably excludes numerous representative oligos, usually from the same members of a gene family, leading to a lack of information for matching the tiling-array signals with gene structure.
4. Positional effect is also a crucial factor that needs to be taken into account because the 3′ end of cDNA is more readily labeled than the 5′ end during reverse transcription.

Experimental design of the high-resolution rice tiling microarray

To determine the impact of GC composition, melting temperature, and sequence repetitiveness of oligo probes from a given gene on their hybridization strength, we selected a 1-Mb portion (from 10 to 11 Mb) from rice japonica chromosome 10 to design a high-resolution tiling microarray. In this design, 36-mer oligos in 5-bp step were used to tile both strands of the given genomic region (Fig. 1A). Some 194,497 oligo pairs were generated in total. Importantly, as an absolutely unbiased survey, all oligos were retained and synthesized. All the oligos were arranged on slides in a "chessboard" layout, which means that every positive feature (interrogating oligo) was surrounded by four negative features (blank position) for background noise subtraction in further data processing. Hybridization targets were prepared from four major rice tissues—seedling root, seedling shoot, panicle, and suspension-cultured cells—labeled with Cy3 dye. Hybridized slides were scanned by a GenePix 4000B scanner (Axon, Foster City, Calif.) using the 532-nm channel and visualized with GenePix Pro 3 image analysis software (Axon).

Tiling-array raw data processing workflow

The raw data were processed in the following order: (1) Positive and negative intensities were extracted separately, and deposited in a uniform format. (2) To smooth the variance between positive and negative intensities, arising from targets' different affinity to DNA and blank linker, a global normalization was assigned to scale positive and negative distribution to a consistent baseline. (3) To screen the background noise, the average intensity of four surrounding negative intensities was calculated and then subtracted from the centric positive intensity. (4) After filtering the negative value, the remaining noise intensities were corrected by using a hypothesis test to

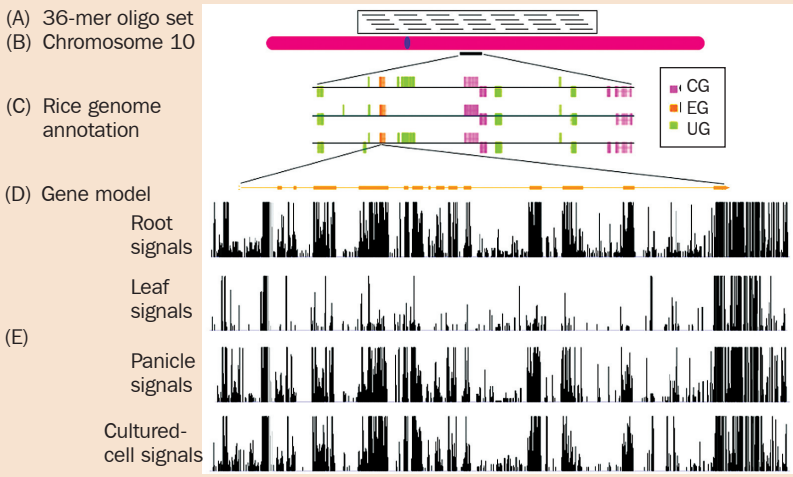

Fig. 1. High-resolution tiling-path microarray was employed to interrogate a 1-Mb region of japonica chromosome 10 (10 to 11 Mb). (A) A 36-mer oligo set was used to tile the given genomic sequence on both strands. Each of two adjacent oligos overlapped 31 bp in length. (B) Schematic representation of chromosome 10. A black bar marks the interrogated region. (C) Zoomed genome annotation set in part of a selected region, and all the protein-coding gene models were classified into three categories: CG = full-length cDNA-confirmed genes; EG = EST-supported genes; UG = predicted genes without any expression support. (D) A candidate EG gene model was selected as an example, denoted by arrowed block, whose point indicates the transcription direction. (E) Tiling-array signal (vertical lines) of four parallel experiments hybridized with different RNA populations pooled from root, leaf, panicle, and suspension-cultured cell.

determine whether a probe represented noise or a signal. The outputs of mature data were then transferred to advanced analysis for mapping signal probes to annotated gene models, scanning for novel transcription units, and further analyzing oligo hybridization properties.

Mapping tiling-array signals to annotated gene models

To measure the detection efficiency of tiling experiments, both annotated gene models and oligos (with their hybridization intensities) were mapped back to the interrogated genomic region according to their chromosomal coordinates. A visualization platform was developed under the PERL environment plus SVG package, to directly assess the matching of gene models and tiling-array signals. In total, 101 BGI annotated protein-coding gene models were localized in the given region, including 36 full-length cDNA- and EST-supported gene models. A majority of the annotated models showed agreement to signal clusters to a certain extent (Fig. 1), and an obvious difference can be observed in that the probe signals in exonic regions were more compactly clustered than in intronic regions and intergenic regions, indicating that a well-designed algo-

rithm could effectively distinguish exons and introns by recognizing characterized signal clusters.

We used the same algorithm to calculate the detection rate of 101 annotated gene models both by a medium-resolution tiling experiment (36-mer oligo with 10-bp interval) and high-resolution tiling experiment (36-mer oligo with 5-bp step). The latter experiment confirmed slightly more gene models than our former analysis (Li et al 2005).

Oligo hybridization properties analysis

In our oligo hybridization properties analysis, we primarily focused on three fundamental factors that influence hybridization experiments according to prior knowledge and experience in analyzing microarray data: oligo GC composition, melting temperature, and repetitiveness. Two single-exon full-length cDNA verified gene models were selected in our initial analysis. As shown in Figure 2, AK106921 has an open reading frame of 2,635 bp in length. Average melting temperature and GC content of the probes tiling AK106921 are 81.7 ^{o}C and 59.7%, respectively. The abundant signal clusters show a high relationship with GC content and Tm value along the transcribed region, except for the two low-GC-content portions in the 3′ and 5′ end (nearly 30%), causing a shortage of hybridization signals. Another full-length cDNA, AK106577 (3,014 bp), shows a low GC composition (37.57% on average), which leads to an overall scarcity of hybridization signals. In contrast, some highly repeated regions, indicated by red peaks (Fig. 2), show a high correlation with signal clusters. These could be false signals caused by cross-hybridization with other transcripts in the genome.

Relationship between GC composition and hybridization strength

One unique property of *Gramineae* genes is compositional gradients in GC content, codon usage, and amino acid usage, along the direction of transcription, beginning at the junction of 5′-UTR and the coding region (Wong et al 2002). As Figure 3A shows, the distribution of oligo GC content sampled from exonic regions has two peaks, with the low GC peak at about 40% and the high GC peak at about 70%. In contrast, the GC composition in intronic regions apparently has no gradient change, and has a single peak at slightly lower than 40%.

High GC content causes unexpected nonspecific hybridization because of increased oligo melting temperature. To estimate this influence, the average hybridization intensity within every GC content range was calculated and plotted (Fig. 3B). From this curve, we can see that hybridization strength was enhanced linearly with the increase in GC content, when GC content is below 63.4%. From 63.4% to 80.5%, hybridization strength reaches a relatively stable status. After this region, from 80.5% to 90.3%, the sharply increased intensities indicate the existence of nonspecific cross-hybridization that affects the hybridization signals. We selected a multiexon full-length cDNA gene model, AK072682, to illustrate the influence of partially high GC composition within genes. Figure 4A shows the trend of GC compositional gradient change in the given

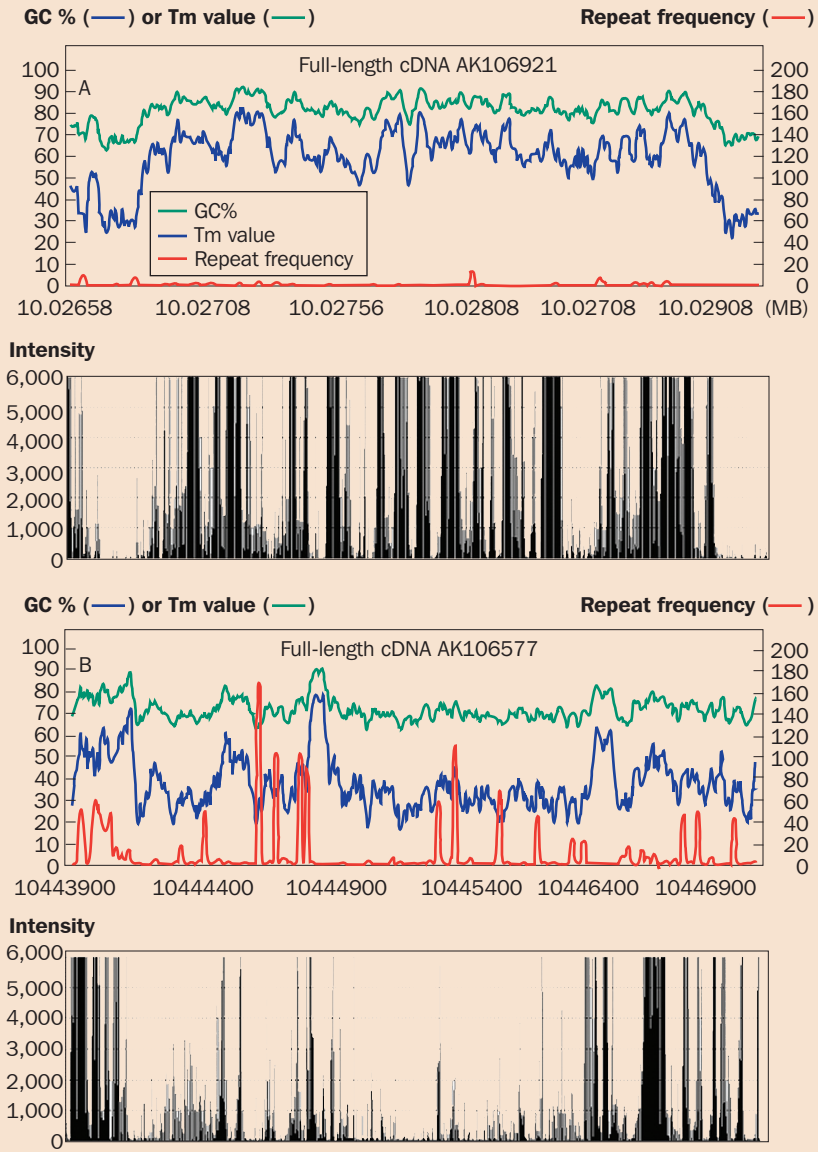

Fig. 2. Overview of three crucial factors that potentially influence the specificity and sensitivity of hybridization experiments: including melting temperature (green curve), GC composition (blue curve), and repetitiveness (red curve) of oligos. Two single-exon full-length cDNA genes were selected to study the relationship among GC%, Tm, repeat, and hybridization strength. (A) AK106921, whose average Tm value and GC% are 81.7 ºC and 59.7%, respectively, has abundant signals except for the two portions of the 3´ end and 5´ end region because of low GC content (30% on average). (B) AK106577, whose average Tm value and GC% are 71.2 ºC and 37.6%, is scarce of hybridization, but some regions show an obvious relationship to a highly repeated feature, as the sharp red peaks indicate.

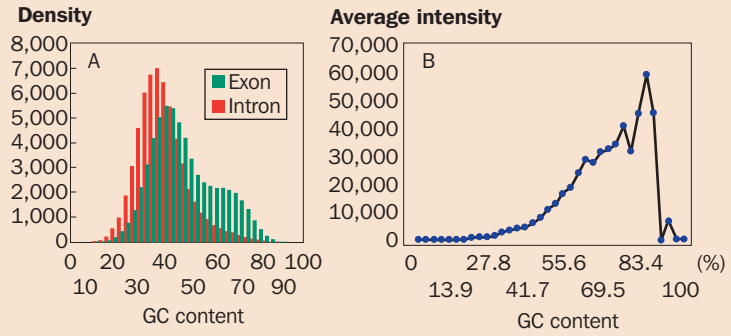

Fig. 3. Influence of oligo GC composition. (A) Distributions of oligo GC content in exonic region (in green) and intronic region (in red). (B) Plotting of GC content against average hybridization intensity.

gene model from transcription start, which is consistent with changes in hybridization intensities (Fig. 4B).

Relationship between oligo melting temperature and hybridization strength

The melting temperature of each 36-mer oligonucleotide was calculated by DAN, a program of the EMBOSS package, based on the nearest-neighbor thermodynamics algorithm. In fact, determination of an appropriate Tm range is crucial for the outcome of hybridization experiments. The Tm value of exonic oligos shows a similar distribution with that of the GC content of exonic oligos, indicating that the potential impact of melting temperature might take effect in the same way as oligo GC composition. In our further study of the impact of melting temperature on hybridization, average hybridization intensity within each Tm region was calculated and plotted, as Figure 5 shows. From this curve, we easily recognized a relatively stable intensity area that corresponds to a Tm range of 82 to 88 ^{o}C, leading us to the conclusion that 85 ±3 ^{o}C might be the optimal temperature for 36-mer oligo-based microarray experiments.

Impact of oligo repetitiveness on hybridization experiments

To evaluate the impact of oligo cross-hybridization, we determined the minimal aligned length and copy number of each probe against all annotated genes in the rice genome. To this end, we blasted all 388,954 oligos against the 45,797 cDNA sequences of rice protein-coding genes released by TIGR annotation. The number of aligned sequences with length ranging from 14 to 35 bp for each oligo provides an estimation of the nonspecific matching with gene models located in other regions of the genome. Surprisingly, the short aligned sequences with lengths of about 15 bp have a relatively high occurrence frequency, which is supposed to be some kind of common repeat motif in a number of genes.

To determine the minimal similarity length that might influence the reliability of hybridization, the average intensity of every similarity length was calculated and

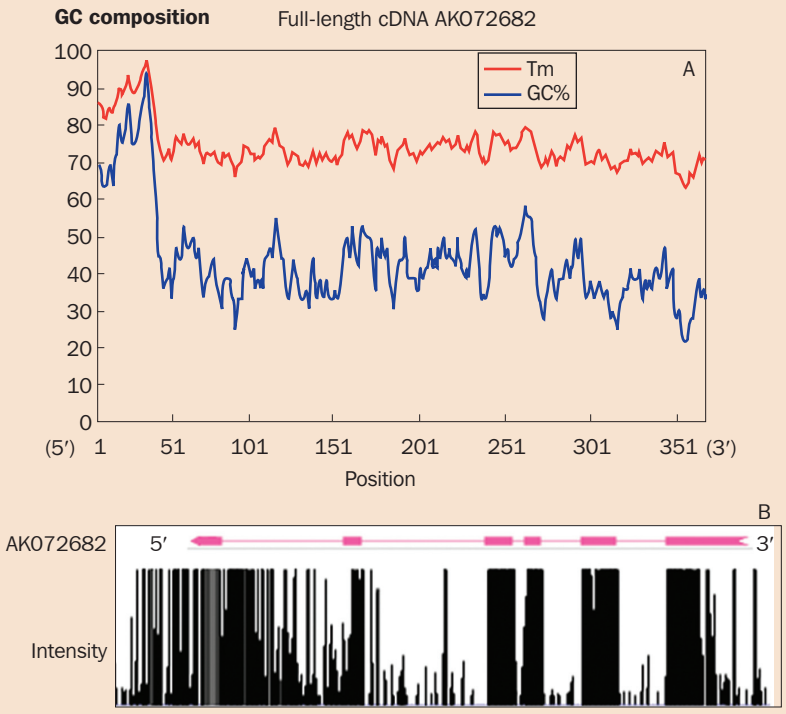

Fig. 4. An example of FLcDNA gene AK072682 to illustrate the positional effect in hybridization experiments, caused by the property of GC compositional gradient in the rice gene. (A) GC content and Tm value gradient change along the transcription direction. (B) The high-GC feature of oligos in the neighborhood of transcription start introduces strongly nonspecific hybridization.

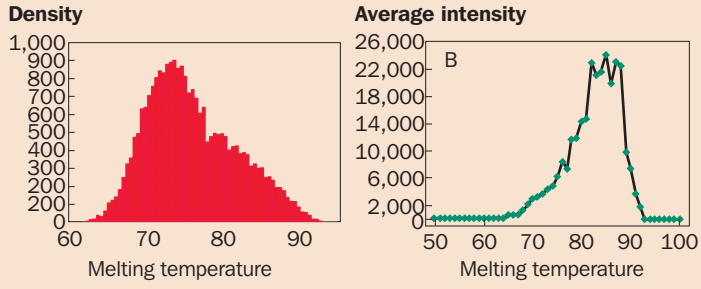

Fig. 5. (A) Distribution of oligo melting temperature in exons shows the identical shape with GC content. (B) Plotting of melting temperature against average hybridization intensity indicates that the relatively stable hybridization strength corresponds to optimal temperature conditions from 82 to 88 °C.

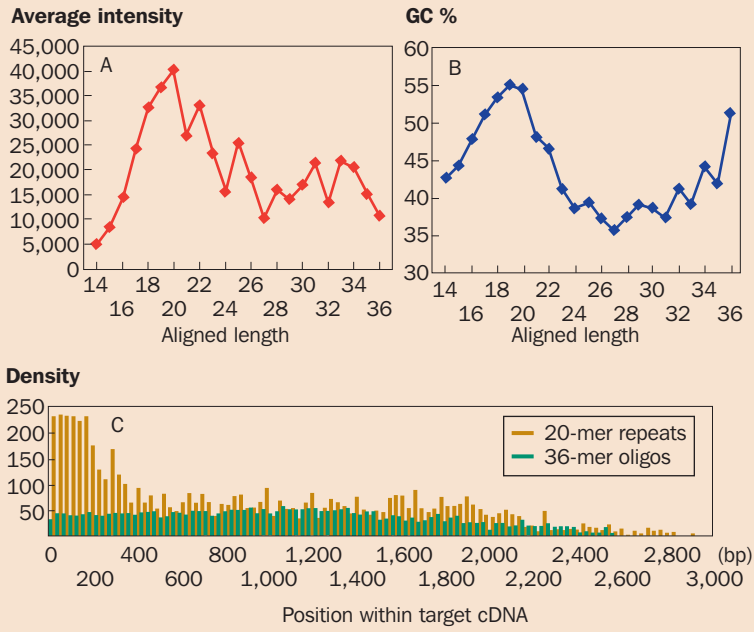

Fig. 6. High GC characterizing 20-mer repeats introduces strong nonspecific cross-hybridization. (A) Plotting of average intensity against different aligned length indicates that around 20-mer repeats have the highest hybridization level. (B) Plotting of average GC content against aligned length indicates that this phenomenon can be interpreted by the high-GC property of the 20-mer repeat. (C) The 20-mer repeat can usually be located within the first 400 bp after the transcription start point. The brown histogram is the distribution of the 20-mer repeat position within all the aligned target cDNA genes. The 36-mer oligos were aligned evenly along the target genes as the green histogram shows.

plotted against them, as Figure 6 shows. Contrary to our expectation, the oligos of around 20-mer similarity length have significantly higher hybridization strength than even 36-mer perfectly matched oligos, designed originally from the interrogated gene models. A further study to survey the relationship of GC composition and oligo similarity can interpret this phenomenon, which is illustrated in Figure 6B. In this plotting of oligo similarity length against average GC content, the 20-mer repeat motif seems to bias toward a higher GC composition than longer repeat motifs above 24 bp. By locating this kind of 20-mer repeat inside the aligned genes, we found that they are more abundant in the 5′ portion of genes than in the middle and 3′ portion of genes, as Figure 6C shows. Therefore, we could conclude that the high GC featured 20-mer repeats and the GC-biased usage of codons in the neighborhood of a gene's transcription start influence the specificity of hybridization experiments and generate most of the false positive signals that partially interfere with verification of gene structure, mostly around the transcription start point.

Conclusions and prospects

The development of tiling-path microarrays provides an efficient way to detect transcriptional activities on a genome scale. In our analysis of a high-resolution rice-tiling array, designed in a 1-Mb region on japonica chromosome 10, we made a series of analyses on three major factors that influence the specificity and efficiency of hybridization: oligo GC content, melting temperature, and repetitiveness. The feature of GC content gradient change from the 5′ end to 3′ end of the rice gene seems to be one of the crucial factors that cause nonspecific and unstable hybridization. Extremely high GC content (above 80%) causes drastically increased cross-hybridization. However, as Tm value ranges become narrower along with an increase in GC content, we identified a relatively stable melting temperature range, from 82 to 88 ^{o}C, for 36-mer oligo-based microarray design. Further analysis of the repetitiveness of all oligos revealed unexpected short repeat motifs around 20 bp in length, having relatively high GC content. These short repetitive motifs could influence the specificity of hybridization and the determination of gene expression level.

References

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. 2004. Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. 7:732-736.

Bertone P, Gerstein M, Snyder M. 2005. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res. 13:259-274.

Hoshikawa K. 1993. Science of the rice plant. Vol 1. Morphology. Tokyo (Japan): Nobunkyo. p 133-186.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436:793-800.

Jiao Y, Jia P, Wang X, Su N, Yu S, Zhang D, Ma L, Feng Q, Jin Z, Li L, et al. 2005. A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. Plant Cell 17:1641-1657.

Li L, Wang X, Xia M, Stolc V, Su N, Peng Z, Tongprasit W, Li S, Wang J, Wang X, Deng XW. 2005. Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. Genome Biol. 6:R52.

Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. 2005. Applications of DNA tiling arrays for whole-genome analysis. Genomics 85:1-15.

Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA, Yu J. 2002. Compositional gradients in Gramineae genes. Genome Res. 12:851-856.

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, et al. 2005. The genomes of *Oryza sativa*: a history of duplications. PLoS Biol. 3:e38.

Notes

Authors' addresses: X. Wang, L. Li, and X.W. Deng, National Institute of Biological Sciences, Zhongguancun Life Science Park, Beijing 102206, China; L. Li, V. Stolc, and X.W. Deng, Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, e-mail: xingwang.deng@yale.edu; X. Wang and S. Li, Peking-Yale Joint Research Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences, Peking University, Beijing 100871, China; X. Wang, C. Chen, J. Wang, and S. Li, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China; V. Stolc, Genome Research Facility, NASA Ames Research Center, MS 239-11, Moffett Field, CA 94035; W. Tongprasit, Eloret Corporation, Sunnyvale, CA 94087.